Education as a Protective Factor: Exploring the Link Between Educational Attainment and Police Stops

Jiaheng Li

Pinehurst School, Auckland, New Zealand

jl050884@pinehurst.school.nz

Abstract. Crime has long been studied from economic and social perspectives, but the direct role of education in shaping criminal outcomes remains underexplored. This study investigates the relationship between educational attainment and the likelihood of being stopped by police for reasons other than minor traffic offenses. The dependent variable (policever) Indicates whether an individual was ever stopped by police in 1980, while the key independent variable is education, measured both as years of schooling and as categorical groups. Covariates include gender, urban residence, marital status, employment, and age. Multiple Linear Regression models were estimated in Stata16 with Ordinary Least Squares (OLS) to assess the effects. The results demonstrate that higher education levels are consistently and significantly associated with a reduced probability of being stopped, while gender, urban status, and marriage also show important associations, and employment and age remain statistically insignificant. These findings emphasize education as a protective factor against criminal justice exposure, although limitations related to dataset scope and self-reported measures highlight the need for further research.

Keywords: Crime rate, OLS, multiple linear regression.

1. Introduction

Crime rate is defined as the number of reported criminal incidents per unit of population within a given time frame, which is a widely used indicator to assess the safety and stability of a society. Education level refers to the highest degree or level of school completed by individuals which is often used to measure human capital. The relationship between crime rates and education levels has not been the subject of longstanding academic inquiry yet as far. In this case, it is crucial to understand the link between crime rates and education.

Considerable research has been applied to exploring the determinants of crime from a variety of socioeconomic perspectives. For example, Niknami investigates how disparities in individual income levels contribute to variations in criminal behavior [1]. The author found that individuals who earn significantly less than their local peers are more likely to commit crimes by using micro-level data, which stress the psychological and material pressures associated with relative deprivation. Except for the income, other scholars have done research in the relationship between age structure and criminal behavior. For instance, Shulman et al. highlights that biological, social, and economic transitions in adolescence and early adulthood significantly influence criminal propensity [2]. Further study such as the work carried out by Papps et al. focuses on labor market variables. This paper suggests that both cyclical and long-term unemployment can lead to increased crime rates. Moreover, McCuddy investigates the social environment as a driver of criminal behavior, showing how peer effects can increase or decrease the likelihood of crime, especially among youth [3].

Despite the large amount of these studies, there remains a deficiency of focused research concern the direct link of educational levels and crime rates. While education is often included as a secondary or control variable, it has not been the central object of large-scale empirical analysis in most of the cited works. This oversight is notable given the plausible mechanisms through which education could influence crime, such as improved employment prospects, enhanced cognitive skills, and stronger social norms. Therefore, this study seeks to address a critical gap in literature by systematically analyzing how variations in education levels across populations correlate with crime rates, using comprehensive data sources.

2. Methodology

2.1. Dataset Preparation

The data set used in this research is drawn from the National Longitudinal Survey of Youth (NLSY) This study extracted the dataset from the National Longitudinal Survey of Youth, comprising 4,488 individual observations. The dataset includes a total of 29 variables.

The dependent variable is policever in the research, which is a numerical variable, if someone was stopped by a police officer in 1980 for reasons other than minor traffic violations, the code is 1, otherwise it is 0. The independent variable represents years of schooling, truncated at 20 years. The covariates urban (coded 1 if urban location, 0 otherwise), sex, married (coded 1 if married, 0 if not), working (coded 1 if working, 0 otherwise) and age (in years). In addition, the covariate age is a string variable, which represents the gender of each individual, be replaced by sex_1 coded 1 if male, 0 otherwise.

2.2. Multiple Linear Regression

Multiple Linear Regression in Stata16 is used to research the linear relationship of variables mentioned. The main idea of Multiple Linear Regression is to explain and predict the variation in a dependent variable by using a linear combination of multiple independent variables [4, 5]. The principle of the model is that coefficients are estimated using the Ordinary Least Squares (OLS) method [6, 7], which finds the set of coefficients that minimizes the sum of squared differences between the observed values and the predicted values.

The equation of regression is:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \tag{1}$$

OLS Estimation Goal:

$$\min_{\beta_0, \beta_1, \ldots, \beta_k} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2}$$

In this regression model, y_i denotes the dependent variable for observation i, and $\hat{y}_i$ presents its predicted value for that observation. For each observation i, x_{ij} represents the j-th independent variable. The parameter β_0 is the intercept, β_j is the slope coefficient associated with x_{ij}, and ε_i is the random error term.

3. Results and Discussion

In the first regression, the result is:

$$policever = -0.0168 \times eudc + 0.204 \times sex\_0 + 0.0345 \times urban - 0.0175 \times working + 0.000315 age - 0.0501 \times married + 0.320 \tag{3}$$

Table 1. Results of first regression

Variable name	Coefficient	p-value
educ	-0.0168	0.000
sex_0	0.204	0.000
urban	0.0345	0.006
working	-0.0175	0.364
age	0.000315	0.898
married	0.0501	0.000

The estimated coefficient for educ (years of schooling) is −0.0168 with a p-value of 0.000 shown in Table 1, significant at the 1% level. Each additional year of schooling is associated with a 1.68 % decrease in the probability of ever being stopped by the police. The coefficient for sex_0 (coded 1 if male, 0 if female) is 0.204 with a p-value of 0.000, significant at the 1% level, suggesting that males are approximately 20.4% more likely to be stopped than females. The coefficient for urban (1 = urban location) is 0.0345 with a p-value of 0.006, showing a statistically significant positive association at the 1% level, with urban residents being 3.45% more likely to have been stopped than non-urban residents. The coefficient for working (1 = employed) is −0.0175 with a p-value of 0.364, indicating no statistically significant relationship between employment status and the probability of being stopped. The coefficient for age (in years) is 0.000315 with a p-value of 0.898, showing no significant age-related effect on the likelihood of being stopped. Finally, the coefficient for marriage (1 = married) is −0.0501 with a p-value of 0.000, showing significantly a negative association at the 1% level; married individuals are 5.01 % less likely to have been stopped.

In the second regression, the result is:

$$policever = -0.0259 \times edu_{12} - 0.0742 \times edu_{13_{15}} - 0.119 \times edu_{16plus} + 0.204 \times sex_0 + 0.0347 \times urban - 0.0512 \times married - 0.0176 working + 0.0000832 age + 0.161 \quad (4)$$

Table 2. Results of second regression

Variable name	Coefficient	p-value
edu_12	-0.0259	0.127
edu_13_15	-0.0742	0.000
edu_16plus	-0.119	0.000
urban	0.0347	0.006
sex_0	0.204	0.000
married	-0.0512	0.000
working	-0.0176	0.362
age	0.0000832	0.973

In the second model, where education is represented by categorical variables, the coefficient for edu_12 (12 years of schooling) is −0.0259 with a p-value of 0.127 shown in Table 2, indicating a negative but statistically insignificant difference compared to the reference group (edu_0_11, 0–11 years). The regression gets −0.0742 as the coefficient of edu_13_15 with a p-value of 0.000, which has the 1% level significance, corresponding to a 7.42 % lower probability of being stopped. The coefficient for edu_16plus (16 or more years) is −0.119 with a p-value of 0.000, also at the 1% level, corresponding to an 11.9 % reduction relative to the reference group. The coefficient for sex_0 (coded 1 if male, 0 if female) is 0.204 with a p-value of 0.000, indicating that males are about 20.4 % more likely to be stopped than females, significant at the 1% level. The coefficient for urban is 0.0347 with a p-value of 0.006, showing a significant positive relationship at the 1% level. The coefficient for working is −0.0176 with a p-value of 0.362, which is not statistically significant. The coefficient for age is 0.0000830 with a p-value of 0.973, showing no significant relationship with stopping probability. Lastly, the coefficient for married is −0.0512 with a p-value of 0.000, indicating a statistically significant negative relationship at the 1% level; married individuals are 5.12 % less likely to have been stopped.

The regression results in differing importance across covariates. Education shows a strong and consistent negative association with police stops, suggesting that schooling enhances social capital, employment opportunities, and adherence to social norms, thereby reducing encounters with law enforcement.

This aligns with established criminological findings that men are both more frequently involved in risky activities and are subject to stricter policing, with Starr, S. B. showing that male defendants in U.S. federal courts receive substantially harsher sentences than comparable female defendants [8].

Urban residence has a modest but significant positive effect, likely reflecting higher crime density, increased police surveillance, or stricter enforcement in cities. Marital status shows a protective role, with married individuals less likely to be stopped, consistent with the idea that family stability fosters stronger integration into conventional social networks.

By contrast, employment and age were statistically insignificant. Employment may not capture broader economic security once education is considered, while age effects may be muted due to the relatively narrow age range of the National Longitudinal Survey of Youth Cohort.

Overall, education and gender stand out as the strongest predictors, while urban residence and marriage exert moderate influence, highlighting the multifaceted social context underlying crime-related outcomes.

4. Conclusion

This study explored the relevance of educational level and the likelihood that being stopped by police for reasons other than minor traffic offenses. Two regression models were estimated: one with education as a continuous variable and another using categorical indicators, controlling for gender, urban residence, employment, age, and marital status.

Across both models, education showed a consistent and statistically significant negative association with police stops. In the continuous specification, each additional year of schooling was linked to a 1.68 % decrease in the probability of being stopped ($p < 0.01$). In the categorical specification, compared to individuals with 0–11 years of schooling, those with 13–15 years had a 7.42% lower probability, and those with 16 or more years had an 11.9 % lower probability, both significant at the 1% level.

Other covariates also revealed notable patterns. Being male (sex_0 = 1) was strongly associated with a higher stop probability—around 20 % more than females—significant in both models. Urban residence was positively related to stop likelihood (about +3.45 pp), while being married was linked to a roughly 5 % lower probability. Employment status and age were not statistically significant in either model.

Several limitations must be acknowledged. Firstly, National Longitudinal Survey of Youth focuses on a specific group of American youth, limiting its promotion to other age groups or backgrounds. Secondly, the dependent variable relies on self-reported data from 1980, which may involve recall bias and fail to capture changes in policing over time. Thirdly, unobserved factors like neighborhood conditions, peer effects, or local policing practices have not been adequately addressed, which increases the likelihood of omitted variable bias.

Future research should expand to updated datasets and broader populations. Despite these limitations, the research findings emphasize that education is a key protective factor in preventing criminal justice exposure, providing important insights for policy and academic research.

References

[1] Niknami S. The effect of relative income on crime: evidence from micro-data. SOFI, Stockholm University; 2012.

[2] Shulman EP, Steinberg LD, Piquero AR. The age-crime curve in adolescence and early adulthood is not due to age differences in economic status. J Youth Adolesc. 2013;42 (6):848-60.

[3] McCuddy T. Peer delinquency among digital natives: The cyber context as a source of peer influence. J Res Crime Delinq. 2020;58 (3):306-42.

[4] Eberly LE. Multiple linear regression. Topics Biostat. 2007 Jan 1;165-87.

[5] Jobson JD. Multiple linear regression. In: Applied multivariate data analysis: Regression and experimental design. New York, NY: Springer New York; 1991. p. 219-98.

[6] Zdaniuk B. Ordinary least-squares (OLS) model. In Encyclopedia of quality of life and well-being research 2014 (pp. 4515-4517). Springer, Dordrecht.

[7] Dismuke C, Lindrooth R. Ordinary least squares. Methods and designs for outcomes research. 2006 Jan;93 (1):93-104.

[8] Starr SB. Estimating gender disparities in federal criminal cases. Am Law Econ Rev. 2015;17 (1):127-59.